VALIDATING SPEAKING TEST RATING SCALES THROUGH MICROANALYSIS OF FLUENCY USING PRAAT

Parvaneh Tavakoli (University of Reading)

Fumiyo Nakatsuhara (CRELLA, University of Bedfordshire)

Ann-Marie Hunter (St. Mary's University)

CRELLA Summer Seminar, 6 July 2017

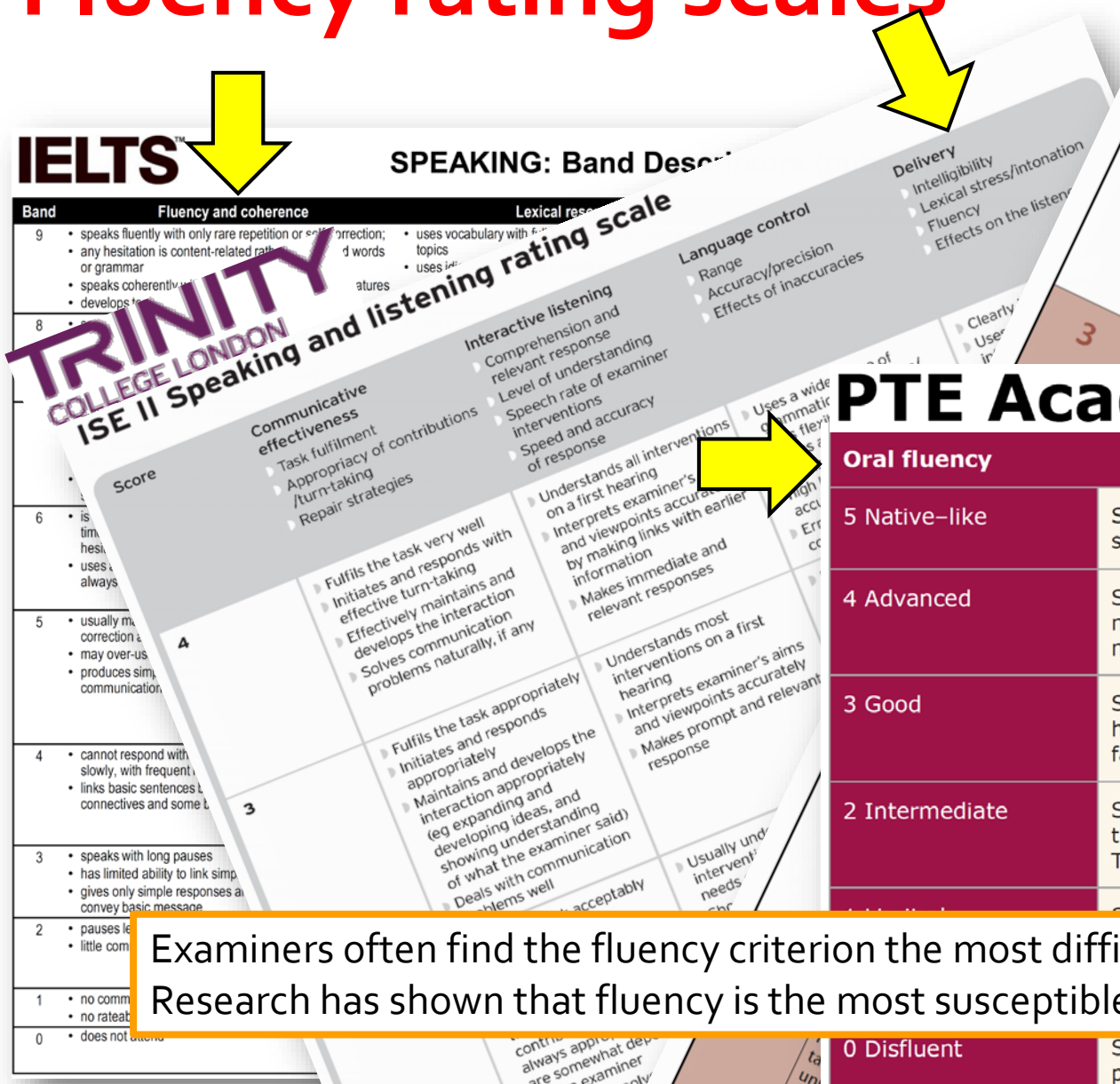
Acknowledgement

- This presentation draws upon a research project funded by the British Council, and carried out under the Assessment Research Awards and Grants programme 2016.

- Any opinions, findings, conclusions or recommendations expressed in this material are those of the presenters and do not necessarily reflect the views of the British Council, its related bodies or its partners.

BACKGROUND

Fluency rating scales



Examiners often find the fluency criterion the most difficult to assess (e.g. Brown 2006b)

Research has shown that fluency is the most susceptible feature to elicitation tasks (e.g. Nakatsuhara 2012)

Approaches to speaking rating scale development/validation

- **Empirical analysis of test-takers' speech samples** (e.g. Brown 2006a; Fulcher 1996; Fulcher, Davidson & Kemp 2011; Galaczi 2013; Nakatsuhara 2014; Turner & Upshur 1995)
- **Raters' perceptions of proficiency when rating spoken performances** (e.g. Brown 2006b; Brown & Ducasse 2009; May 2009; Orr 2002; Pollitt and Murray 1996)
- **Measurement-driven approach as embodied in the CEFR** (e.g. North and Schneider 1998)

Fluency research in SLA and Language Assessment

Certain aspects of fluency are good predictors of oral proficiency:

- Speed fluency (De Jong et al. 2012)
- Speed fluency and number of filled pauses (Revesz et al. 2014)
- Speech rate and mean length of run (Inoue 2013; Kahng 2014)

Fluency is task dependent:

- Impact of information structure (e.g. Tavakoli & Skehan 2005); storyline complexity (e.g. Tavakoli & Foster 2008); intentional reasoning (e.g. Ishikawa 2008); planning time (e.g. Wigglesworth and Elder 2010); discourse mode (e.g. McCarthy 2010; Tavakoli 2016)

Aptis (General) Speaking Test

Aptis

- A **quick**, **flexible and affordable** English language proficiency test designed to meet the diverse needs of organisations and individuals around the world
- **A non-certificated test:** Designed to offer users an alternative to currently available high-stakes certificated examinations.
- It assesses test-takers from **A2 to C1** (O'Sullivan 2015)

Aptis Speaking

Part	Task	Target Level	Rating Scale	Response Time
1	Respond to 3 questions on personal topics	A1/ A2	Scale for Task 1	30 secs x 3
2	Respond to 3 questions, inc. describing a photo and answering a concrete familiar topic related to the photo	B1	Scale for Tasks 2&3	45 secs x 3
3	Respond to 3 questions related to 2 contrasting pictures	B1		45 secs x 3
4	Providing a long turn, integrating responses to a set of 3 questions	B2	Scale for Task 4	(1 min prep +) 2 mins

Aptis Speaking Rating Scales

(O'Sullivan & Dunlea 2016)

- Holistic
- Task-specific

Speaking Task 4

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, fluency and cohesion.

6 C2	Likely to be above C1 level.
5 C1	Response addresses all **three** questions and is well-structured. • Uses a range of complex grammar constructions accurately. Some minor errors occur but do not impede understanding. • Uses a range of vocabulary to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices. • Pronunciation is clearly intelligible. • Backtracking and reformulations do not fully interrupt the flow of speech. • A range of cohesive devices are used to clearly indicate the links between ideas.
4 B2.2	Responses to all **three** questions are on topic and show the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding. • Some pausing while searching for vocabulary but this does not put a strain on the listener. • A limited number of cohesive devices are used to indicate the links between ideas.
3 B2.1	Responses to **two** questions are on topic and show the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding. • Some pausing while searching for vocabulary but this does not put a strain on the listener. • A limited number of cohesive devices are used to indicate the links between ideas.
2 B1.2	Responses to all **three** questions are on topic and show the fol • Control of simple grammatical structures. Errors occur wh • Sufficient range and control of vocabulary for the task. Err thoughts. • Pronunciation is intelligible but inappropriate mispronuncia listener. • Some pausing, false starts and reformulations. • Uses only simple cohesive devices. Links between ideas a
1 B1.1	Responses to **two** questions are on topic and show the followir • Control of simple grammatical structures. Errors occur wh • Sufficient range and control of vocabulary for the task. Err thoughts. • Pronunciation is intelligible but inappropriate mispronuncia listener. • Some pausing, false starts and reformulations. • Cohesion between ideas is limited. Responses tend to be
0 A2.2	Response to **one** question is on topic and shows the following • Limitations in vocabulary make it difficult to deal fully with the task. • Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener. • Noticeable pausing, false starts, reformulations and repetition.

Balance between 'Construct Coverage' and 'Rater Usability'

(relatively long & detailed) (relatively short & succinct)

(Galaczi et al., 2012)

Research Questions

RQ1: How are various aspects of fluency presented across different **levels** of proficiency (A2, B1, B2, and C1) in the Aptis Speaking test?

RQ2: To what extent is test-takers' fluency affected by **task** design (task type, discourse type and target level)?

METHODOLOGY

Materials

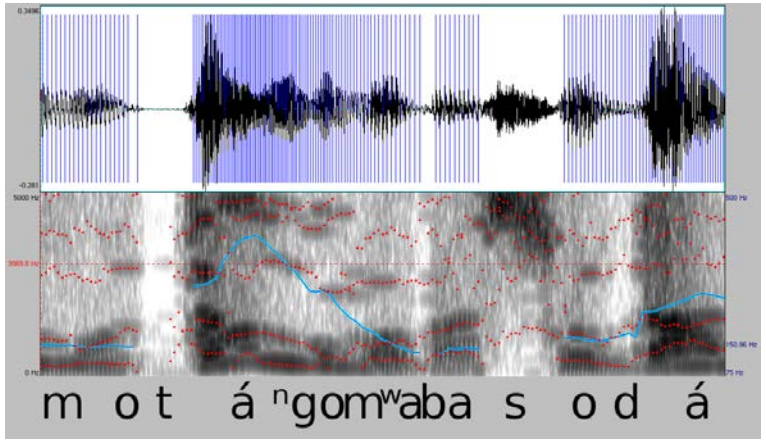
- **32 test-takers in total:** 8 test-takers who were awarded overall scores of A2, B1, B2 and C1

8 test-takers x 4 proficiency levels x 4 tasks = 128 task performances

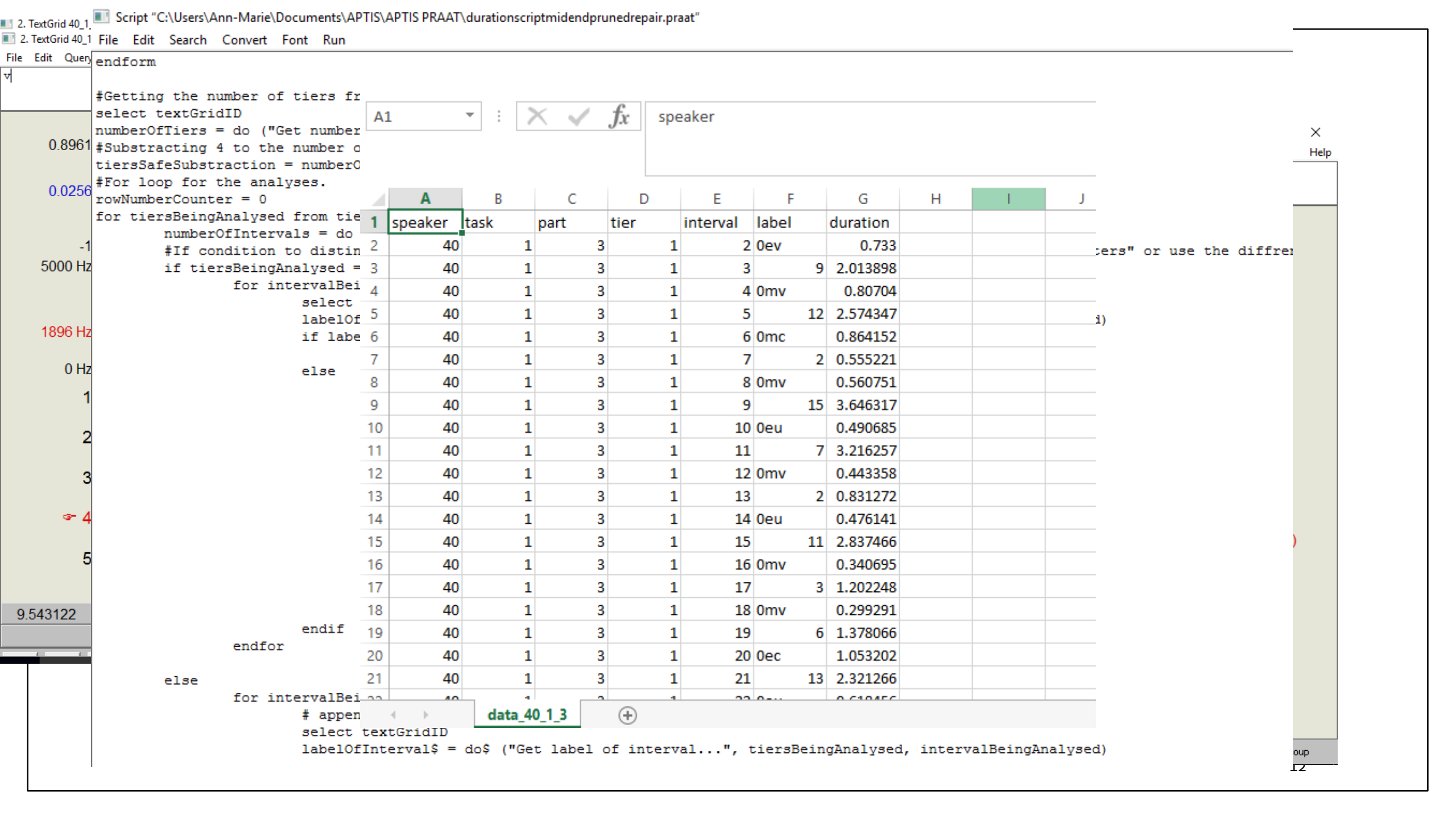
→ 120 performances (without Task 4 performances by A2 candidates)

- An experienced Aptis Speaking rater carefully selected the recordings of the test-takers whose **overall, holistic scores represent their fluency scores** across all 4 tasks.
- **Jagged-profile test-takers** across different components (e.g. Lexis, Grammar) of the holistic scales were **avoided**.

Speech Analysis - PRAAT



- Developed by Paul Boersma and David Weenink at University of Amsterdam primarily for research in the area of phonetics
- Now used in L2 fluency research
  - Allows researchers to study spectrogram
  - 'textgrid to silences' feature (ability to detect and measure silence in a speech sample)
  - Syllable nuclei detection
  - Manual annotation – automatic measurement to extract



	A	B	C	D	E	F	G	H	I	J
1	speaker	task	part	tier	interval	label	duration			
2	40	1	3	1	2	0ev	0.733			
3	40	1	3	1	3	9	2.013898			
4	40	1	3	1	4	0mv	0.80704			
5	40	1	3	1	5	12	2.574347			
6	40	1	3	1	6	0mc	0.864152			
7	40	1	3	1	7	2	0.555221			
8	40	1	3	1	8	0mv	0.560751			
9	40	1	3	1	9	15	3.646317			
10	40	1	3	1	10	0eu	0.490685			
11	40	1	3	1	11	7	3.216257			
12	40	1	3	1	12	0mv	0.443358			
13	40	1	3	1	13	2	0.831272			
14	40	1	3	1	14	0eu	0.476141			
15	40	1	3	1	15	11	2.837466			
16	40	1	3	1	16	0mv	0.340695			
17	40	1	3	1	17	3	1.202248			
18	40	1	3	1	18	0mv	0.299291			
19	40	1	3	1	19	6	1.378066			
20	40	1	3	1	20	0ec	1.053202			
21	40	1	3	1	21	13	2.321266			

data_40_1_3

Fluency measures

Speed measures

a) **Speech rate (pruned):** total number of syllables divided by total performance time (including pauses)

b) **Mean length of run (pruned):** the mean number of syllables between two pauses

c) **Articulation rate (pruned):** total number of syllables per minute divided by total amount of speaking time (excluding pauses)

Breakdown measures

d) **Phonation time ratio:** time spent speaking (excluding pauses)

Length of pauses

e) **Mean length of silent pauses at mid-clause *(e-1)* and end-clause *(e-2)* positions**

f) **Mean length of filled pauses at mid-clause *(f-1)* and end-clause *(f-2)* positions**

g) **Mean length of all pauses (silent AND filled)**

Frequency of pauses

h) **Frequency of silent pauses (per 60 seconds) at mid-clause *(h-1)* and end-clause *(h-2)* positions**

i) **Frequency of filled pauses (per 60 seconds) at mid-clause *(i-1)* and end-clause *(i-2)* positions**

j) **Frequency of all pauses at mid-clause (*j-1*) and end-clause (*j-2*) positions (silent AND filled)**

Repair measures

k) **Frequency of repairs (per 60 seconds)**

l) **Frequency of false starts and reformulations (per 60 seconds)**

m) **Frequency of partial or complete repetitions (per 60 seconds)**

n) **Frequency of self-corrections (per 60 seconds)**

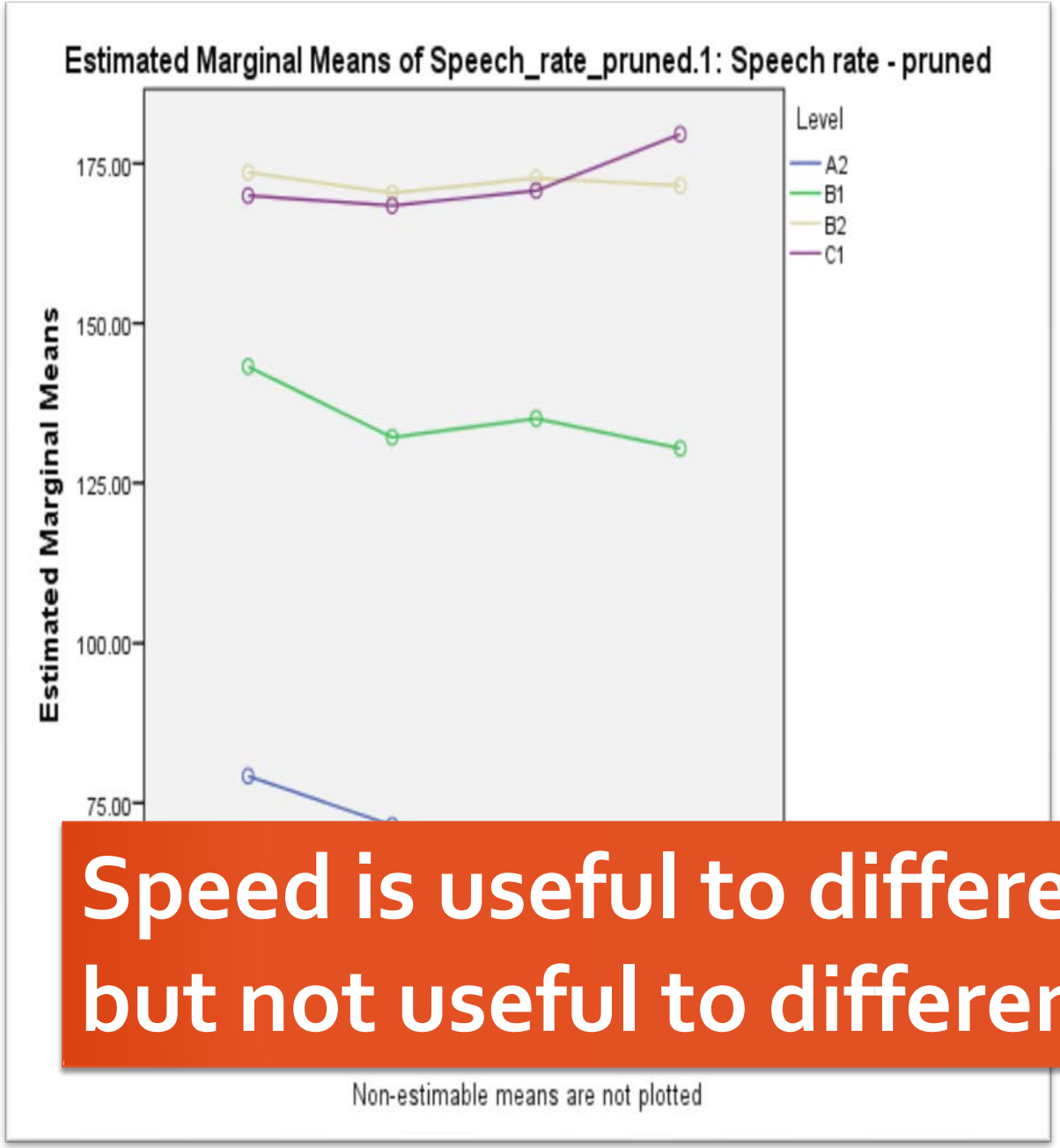
Statistical analysis

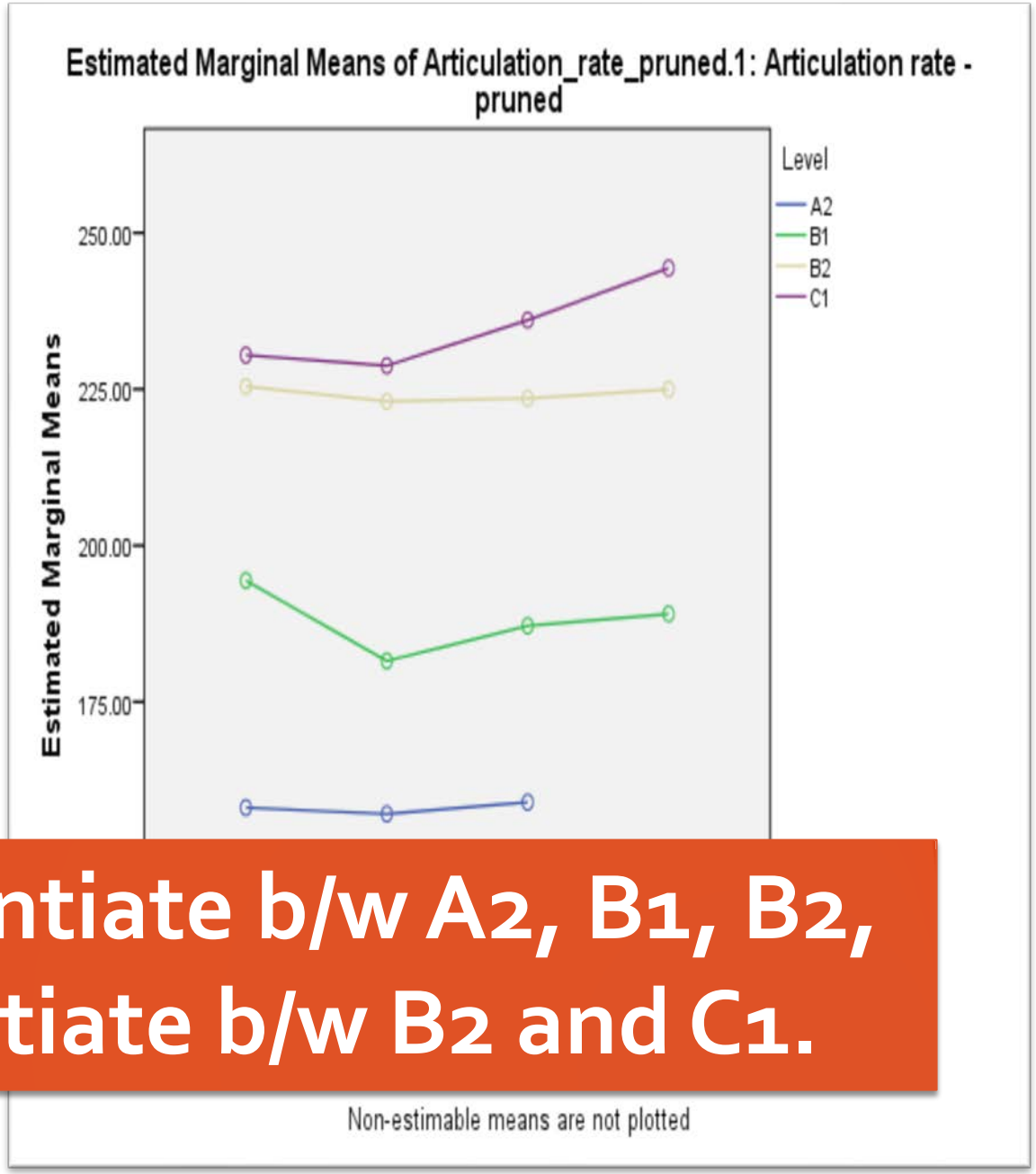
- **A repeated-measures MANOVA**
  - Task (within-participant)
  - Level of proficiency (between-participant)
- **Measures used in RM MANOVA**
  - Speech rate pruned, Number of mid-clause and end-clause pauses per minute, Mean length of mid-clause and end-clause pauses per minute, and a composite repair measure.
- **Results: MANOVA**
  - Proficiency Level (Wilks' Lambda= .160; F= 3.32, p= .000; η^2=.457)
  - Task (Wilks' Lambda= .280; F= 3.63, p= .008; η^2=.720)
  - Proficiency Level and Task (Wilks' Lambda=.097; F= 1.70, p= .04; η^2=.540)
- **Univariate analyses**

[N.B. small sample size + a large number of multiple comparisons → Bonferroni corrections were not used to avoid being too conservative: Results to be interpreted with caution]

ANALYSIS & FINDINGS: LEVELS (RQ1)

(1) Speed fluency distinguishes A2, B1 and B2 levels, but B2 and C1 levels are not different.

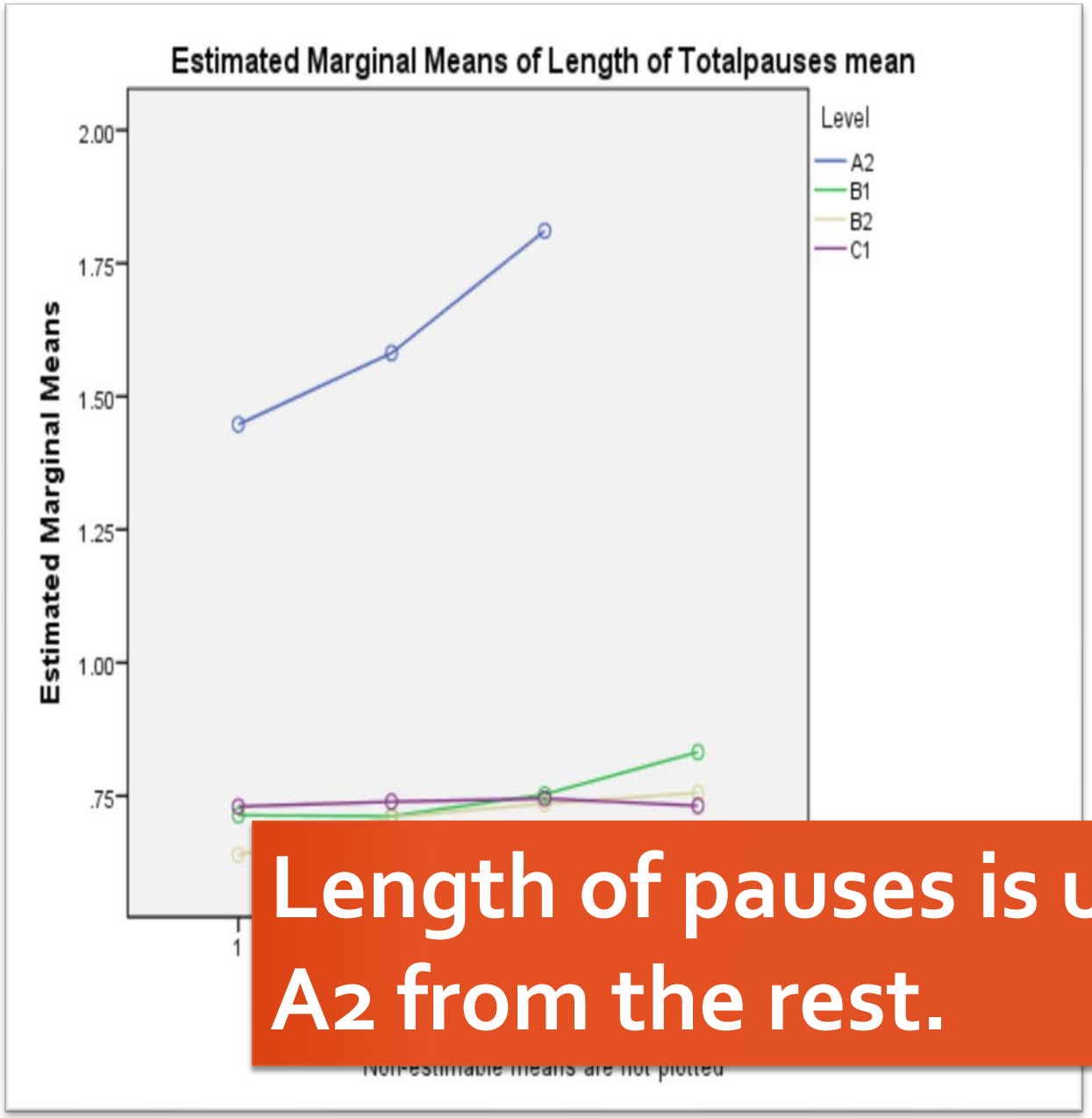


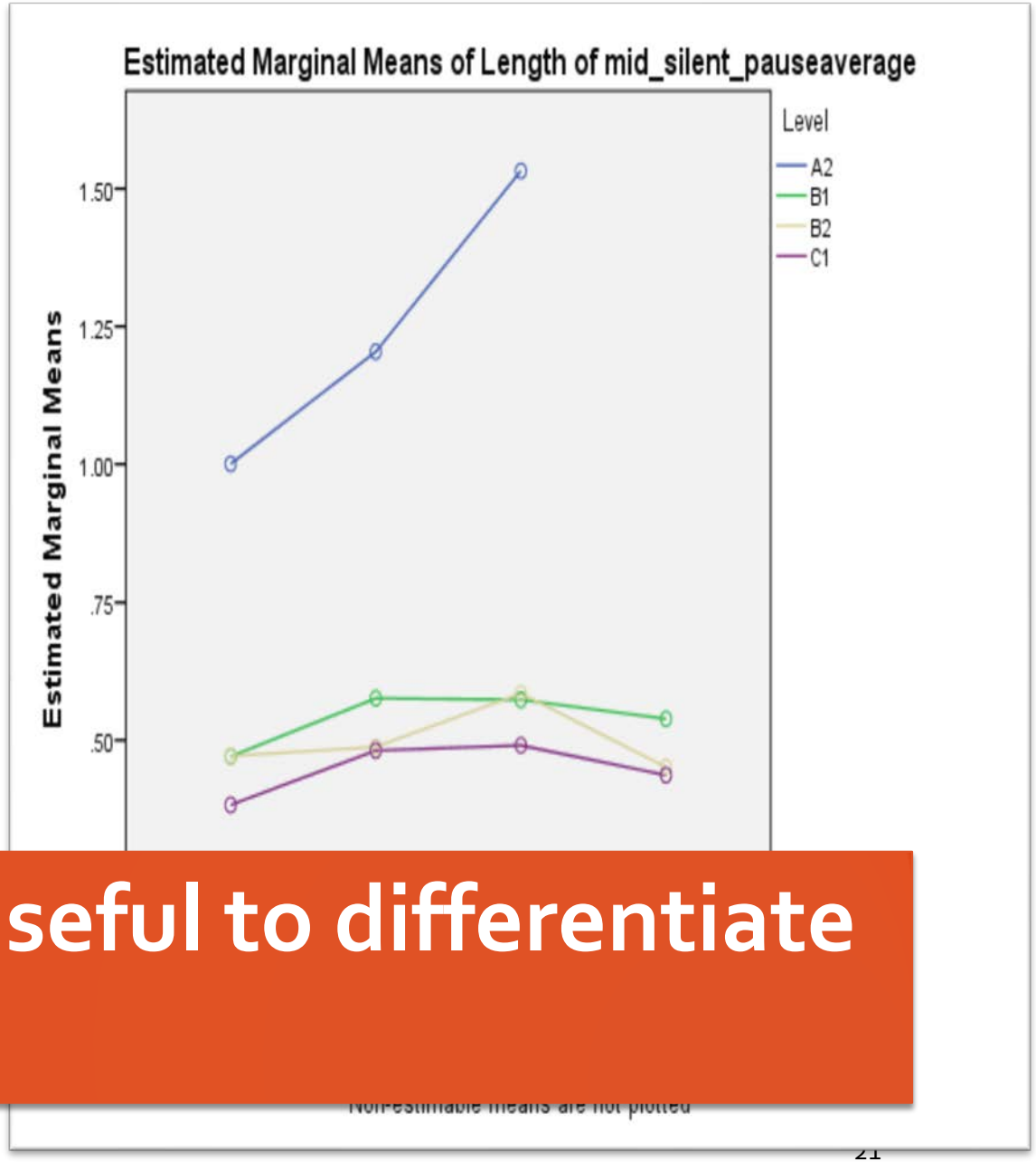


Speed is useful to differentiate b/w A2, B1, B2, but not useful to differentiate b/w B2 and C1.

(2) Length of silent pauses distinguishes A2 level from other levels.

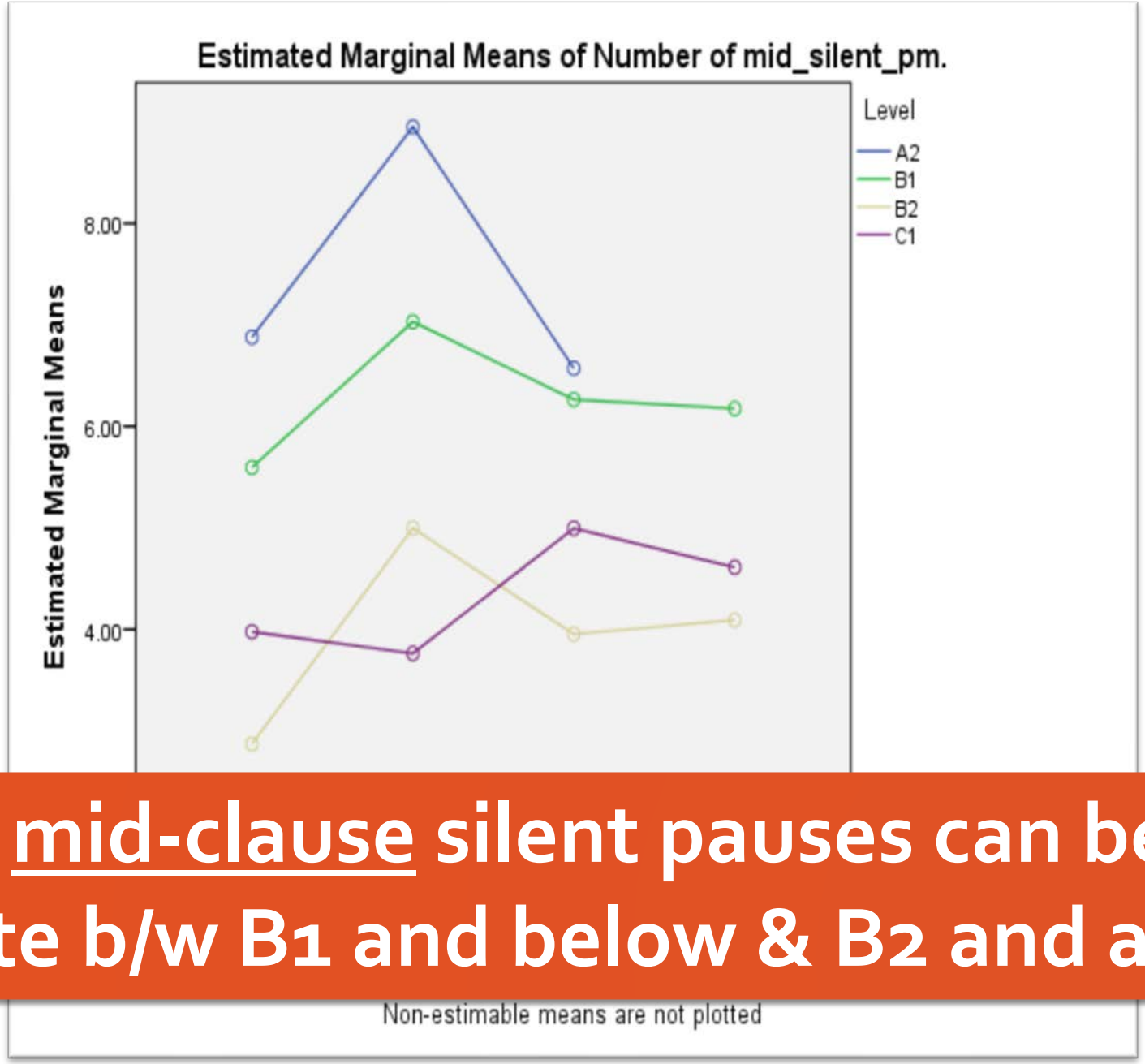
A2 candidates pause much longer than the rest.





Length of pauses is useful to differentiate A2 from the rest.

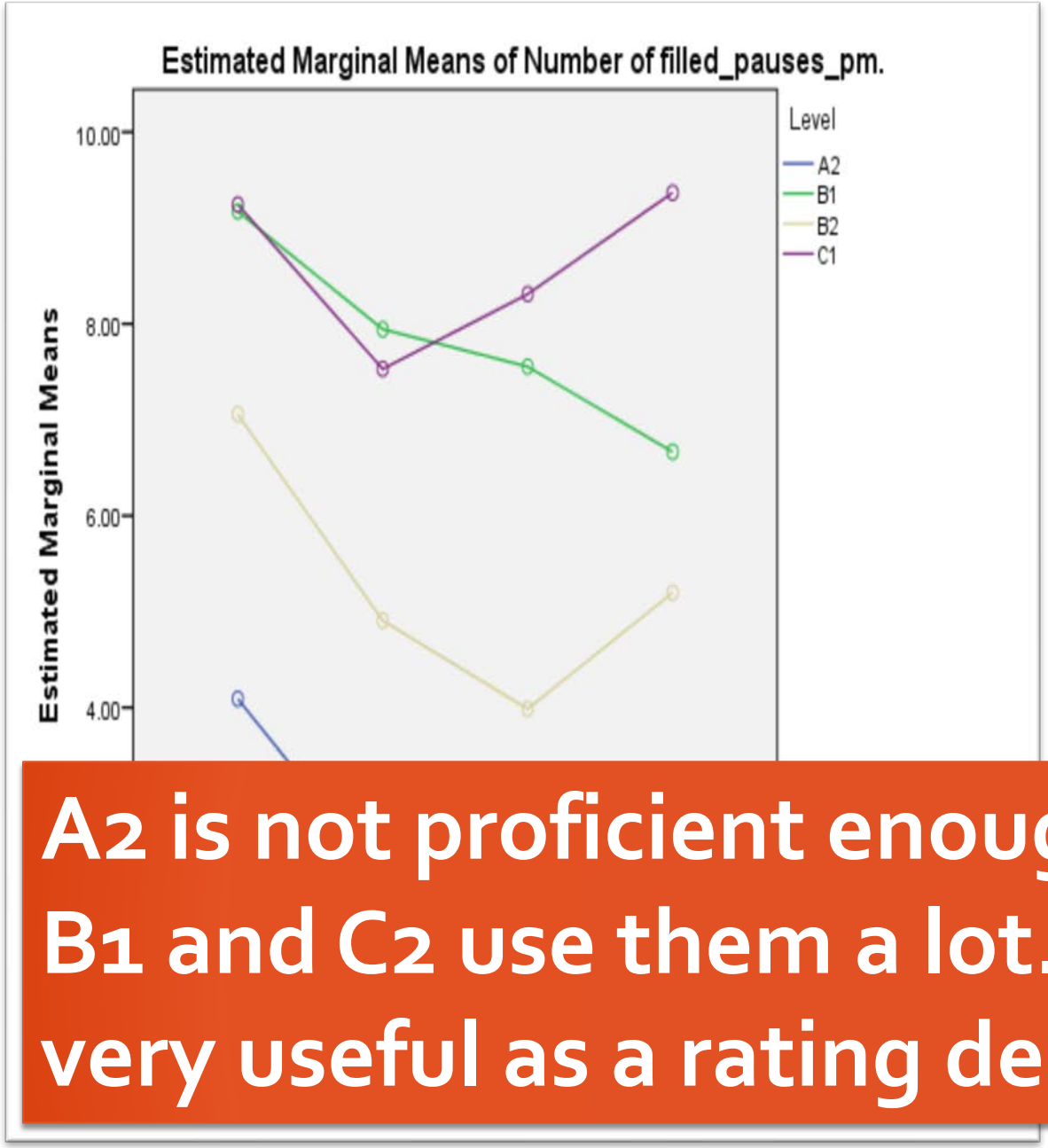
(3) Frequency of mid-clause silent pauses distinguishes lower (A2 and B1) from higher (B2 and C1) proficiency levels.

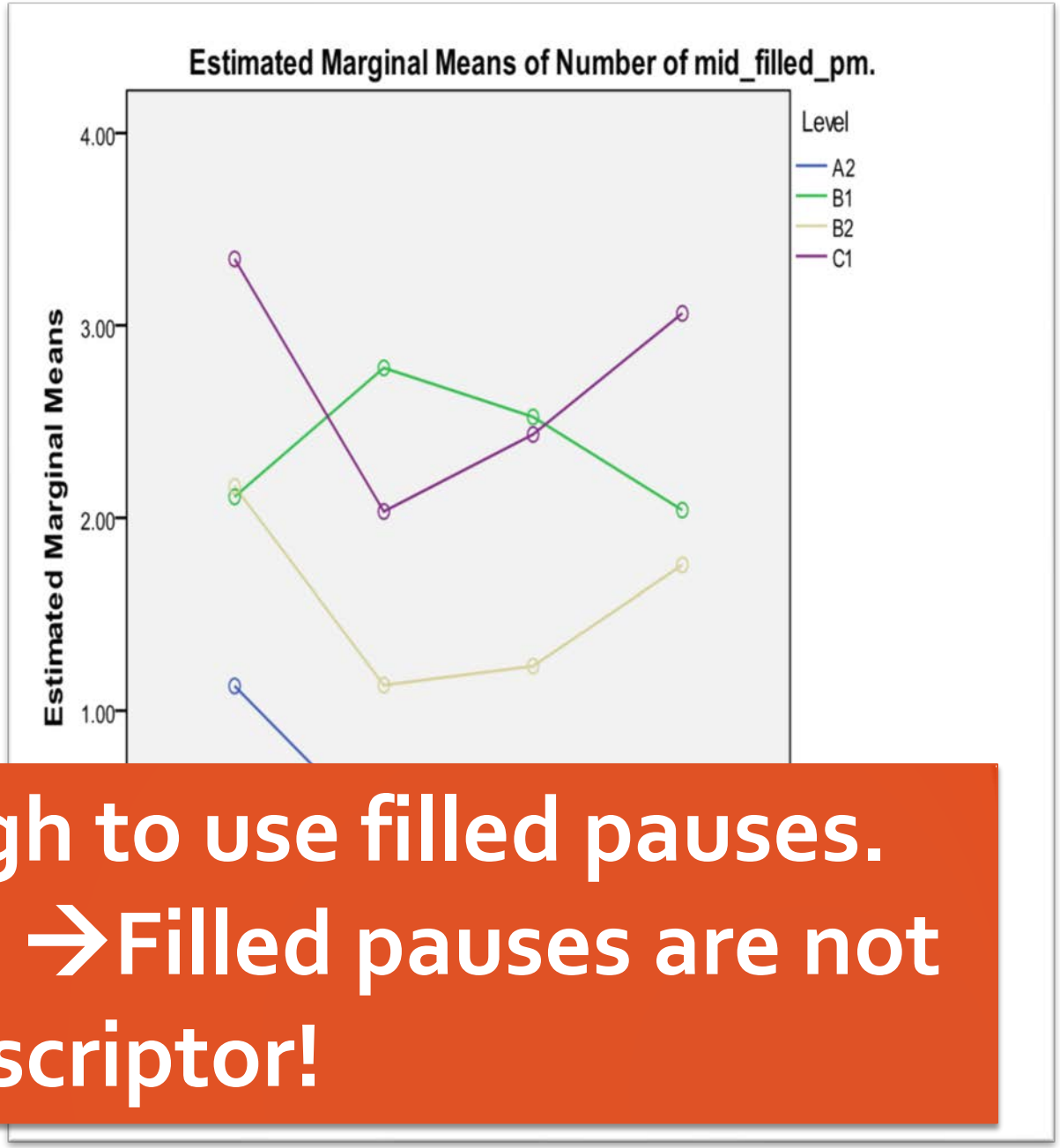


Number of <u>mid-clause</u> silent pauses can be used to differentiate b/w B1 and below & B2 and above.

(4) Frequency of filled pauses distinguishes A2 from higher levels.

B1 and C1 levels use filled pauses more frequently.

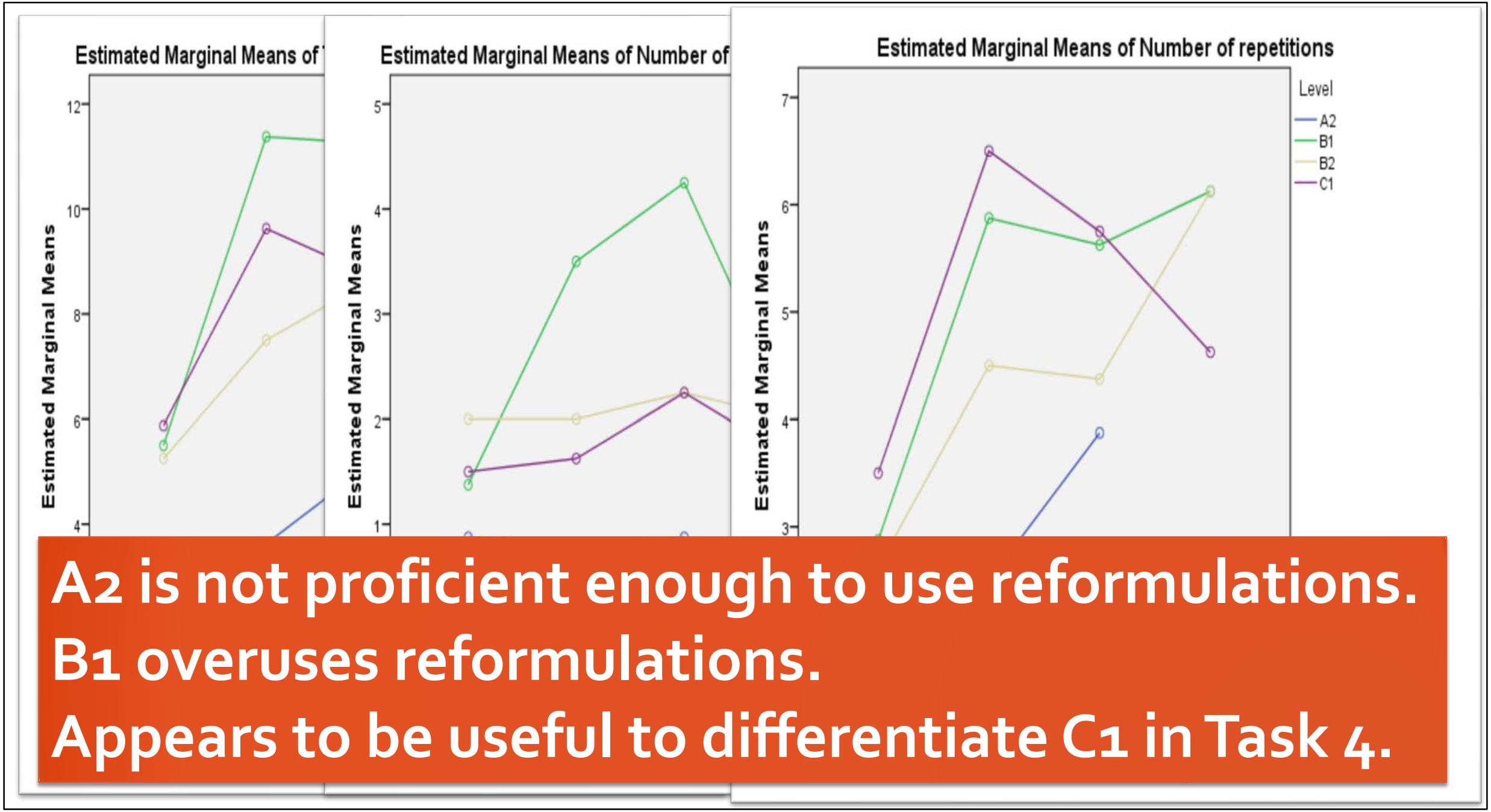




A2 is not proficient enough to use filled pauses. B1 and C2 use them a lot. →Filled pauses are not very useful as a rating descriptor!

(5) Repair measures distinguish A2 and B1 levels; A2 produces very few and B1 most repairs.

B2 and C1 levels use them to a moderated degree.



A2 is not proficient enough to use reformulations.
B1 overuses reformulations.
Appears to be useful to differentiate C1 in Task 4.

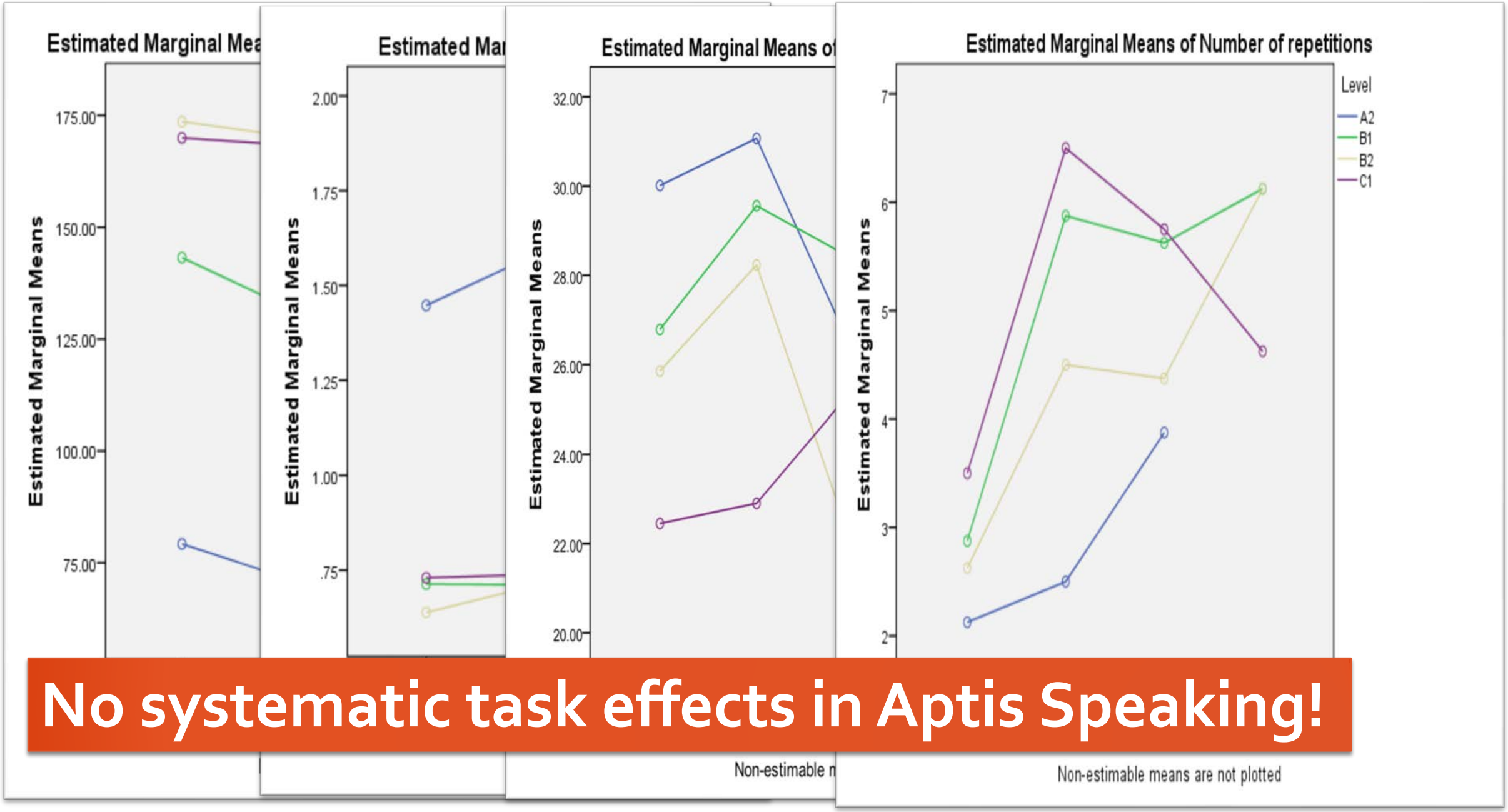
ANALYSIS & FINDINGS: TASKS (RQ2)

(6) Speed of performance is not affected by task type.

(7) Length of pauses is not affected by task.

(8) Frequency of pauses is not affected by task type.

(9) Repair measures distinguish Task 3 from Task 1. Task 3 elicits most repairs.



No systematic task effects in Aptis Speaking!

RECOMMENDATIONS

Modifying the APTIS rating descriptors

Task 1 – fluency rating descriptors

5 B1 (or above)	**Current**	Likely to be above A2 level.
4 A2.2	**Current**	Frequent pausing, false starts and reformulations but meaning is still clear.
	Modified	**Slow speed of speech and long silent pauses but meaning is still clear.**
3 A2.1	**Current**	Frequent pausing, false starts and reformulations but meaning is still clear.
	Modified	**Slow speed of speech and long silent pauses but meaning is still clear.**
2 A1.2	**Current**	Frequent pausing, false starts and reformulations impede understanding.
	Modified	**Slow speed of speech and long silent pauses impede understanding.**
1 A1.1	**Current**	Frequent pausing, false starts and reformulations impede understanding.
	Modified	**Slow speed of speech and long silent pauses impede understanding.**
0 A0	**Current**	No meaningful language or all responses are completely off-topic (e.g. memorised script, guessing).

Tasks 2&3 – fluency rating descriptors

5 **B2 (or above)**	**Current**	Likely to be above B1 level.
4 **B1.2**	**Current**	Some pausing, false starts and reformulations.
	Modified	**Moderate speed of speech but interrupted by mid-clause pauses and reformulations.**
3 **B1.1**	**Current**	Some pausing, false starts and reformulations.
	Modified	**Moderate speed of speech but interrupted by mid-clause pauses and reformulations.**
2 **A2.2**	**Current**	Noticeable pausing, false starts and reformulations.
	Modified	**Slow speed of speech and long silent pauses.**
1 **A2.1**	**Current**	Noticeable pausing, false starts and reformulations.
	Modified	**Slow speed of speech and long silent pauses.**
0	**Current**	Performance below A2, or no meaningful language or the responses are completely off-topic (e.g. memorised script, guessing)

Task 4 – fluency rating descriptors

5 C1	**Current**	Backtracking and reformulations do not fully interrupt the flow of speech.
	Modified	**Natural speed of speech, with some filled pauses and reformulations used effectively.**
4 B2.2	**Current**	Some pausing while searching for vocabulary but this does not put a strain on the listener.
	Modified	**Natural speed of speech, with some pauses and reformulations that do not interrupt the flow.**
3 B2.1	**Current**	Some pausing while searching for vocabulary but this does not put a strain on the listener.
	Modified	**Natural speed of speech, with some pauses and reformulations that do not interrupt the flow.**
2 B1.2	**Current**	Noticeable pausing, false starts, reformulations and repetition.
	Modified	**Moderate speed of speech but interrupted by mid-clause pauses and reformulations.**
1 B1.1	**Current**	Noticeable pausing, false starts, reformulations and repetition.
	Modified	**Moderate speed of speech but interrupted by mid-clause pauses and reformulations.**
0 A1/A2	**Current**	Performance not sufficient for B1, or no meaningful language, or the responses are completely off-topic (memorised or guessing).

CONCLUSIONS

RQ1: Fluency features across different proficiency levels

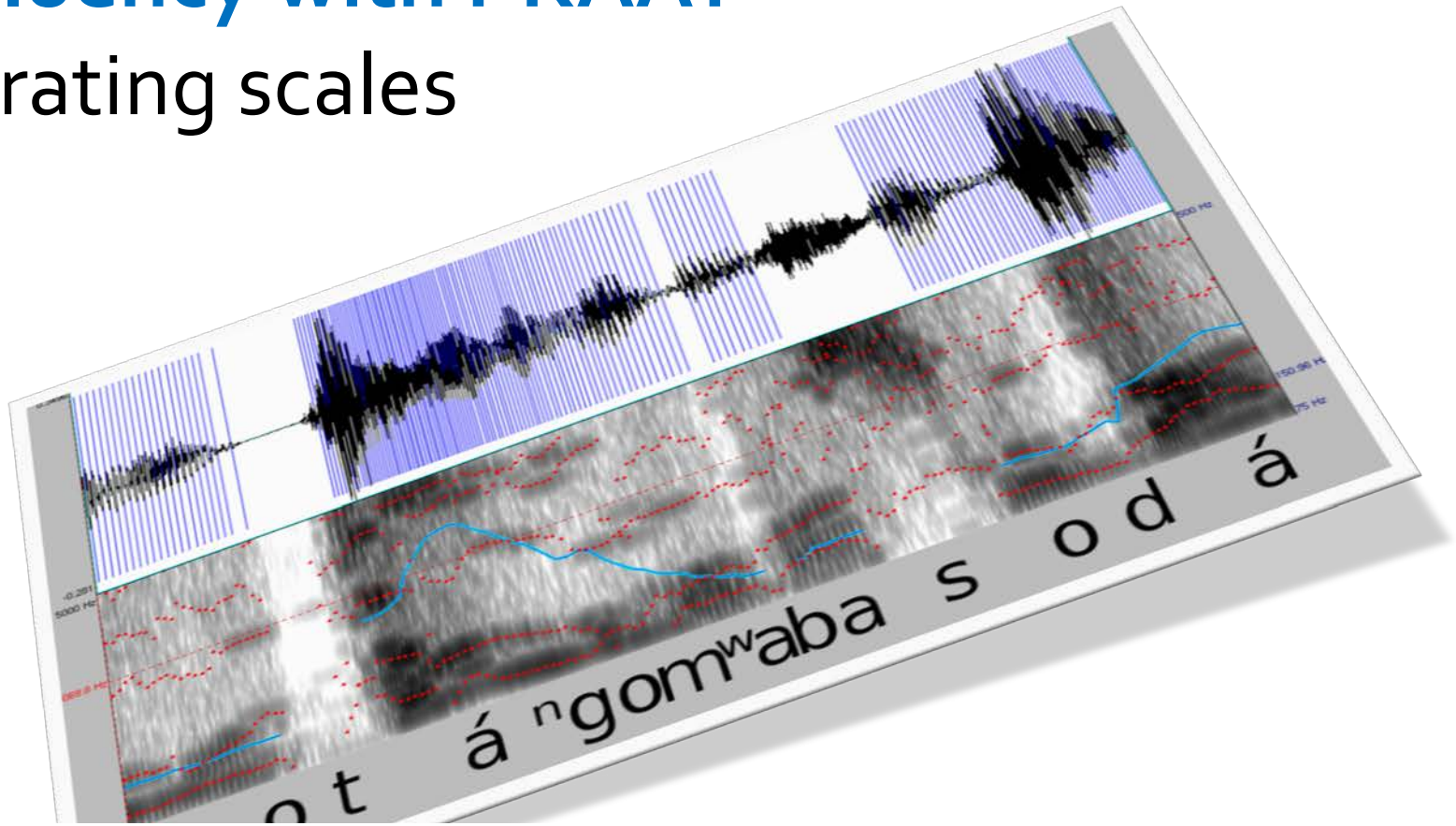
- Some fluency characteristics as **criterial features of A2, B1 and B2/C1** → Can be used to **enhance the scoring validity** of Aptis Speaking
- But, a concern about the **difficulty in differentiating B2 and C1**
  - **Ceiling effect** which comes into play at the B2 level?
  - **Lack of a C1 task**? Not pushing B2 and C1 candidates to their linguistic limit?

RQ2: Fluency features across different tasks

- Surprisingly, **no systematic task effect** ← Perhaps, 4 Aptis tasks are **not distinctive enough**
- This does **NOT invalidate** the Aptis Speaking test or its by-part rating system; **By-part rating system** makes it easier for the examiners to focus on **narrower boundaries** in making judgements.

Micro-analysis of fluency with PRAAT

to validate fluency rating scales



Next step…

Check the extent to which these empirically-informed features are actually salient to human raters in real time!

THANK YOU!

Parvaneh Tavakoli (University of Reading)

Fumiyo Nakatsuhara (CRELLA, University of Bedfordshire)

Ann-Marie Hunter (St. Mary's University)